第 37 卷第 3 期
录用定稿
计算机应用研究
Application Research of Computers
Vol. 37 No. 3
Accepted Paper

基于递归投影的结构性网络嵌入*

孟亚文，傅洛伊，王新兵+

(上海交通大学 IOT 智能物联网实验室, 上海 200240)

摘　要：近些年来，在网络嵌入(network embedding)领域的大多数研究都着眼于基于网络节点邻接关系的社区身份，如 Node2Vec 和 DeepWalk；而基于网络拓扑结构的结构身份的研究则十分匮乏，前沿方法如 struc2vec 等，通常效率很低。提出了 RSNE(recurrent structural network embedding，递归结构性网络嵌入)，一种新颖而高效的结构特征学习方法。RSNE 递归式地把节点的结构身份定义为其邻居结构身份的非线性投影。为了避免退化为基于邻接关系的聚类，采用了一种有效而鲁棒的初始化方法。理论分析显示 RSNE 在时间复杂度上显著优于现有的结构性网络嵌入方法，可视化与量化实验结果也表明 RSNE 在在分类准确性和鲁棒性上达到了最新方法相同或更好的效果，同时消耗的计算时间与空间消耗也远远更少。

关键词：网络嵌入；结构身份；特征学习

中图分类号：TP393.　　doi: 10.3969/j.issn.1001-3695.2018.09.0639

Structural network embedding based on recurrent projection

Meng Yawen, Fu Luoyi, Wang Xinbing+

(*Research Center of Intelligent Internet of Things*, *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*)

Abstract: In recent years, most researches in network embedding, such as Node2Vec and DeepWalk, are focused on the community identity defined by nodes' adjacency, instead of the structural identity defined by topology structure. And state-of-the-art methods of the latter, like struc2vec, are usually time-inefficient. This work presents RSNE(recurrent structural network embedding) , a novel and efficient method to learn node representation from structural identity. RSNE defines a node's structural identity as the non-linear projection of its neighbors' structural identities in a recurrent manner. In order to avoid degradation into clustering with nodes' adjacency, an accurate and robust initialization method based on degrees is applied in this work. Theoretical analysis shows that proposed method is significantly better than the existing methods in terms of time complexity, and can also effectively use hard disk space for memory optimization. Numerical visualized and quantified experiment results suggest that RSNE has equal or better performance than state-of-the-art methods in classification accuracy and robustness while consuming much less computation and time.

Key words: network embedding; structural identity; feature learning

0　引言

目前，人类生活中存在着无数的网络可以用来挖掘出宝贵的信息与知识。例如，学术引用网络可以挖掘出最优秀的文章和最热门的话题，以便学者进行科研。然而，由于网络数据形式复杂，具有计算复杂度高、并行性低、不适用机器学习算法等三个特性 [1]，因此难以直接进行挖掘和学习。近些年来，随着随机游走[2]和词向量嵌入模型(word2vec)[3]在数据挖掘中的引入，网络挖掘领域出现了许多优秀的研究成果，致力于把网络中的节点表示在低维空间中，同时保留其网络结构[1]。最前沿的方法，如 DeepWalk [2]，LINE[4]，node2vec[5]，在节点分类和边预测任务上取得了显著的成果。同时也有研究表明上述提到的方法本质上都能等效为一种封闭形式的特殊矩阵分解[6]。 然而，除了上述方法，最近两年(2017—2018 年)还涌现出了许多崭新的网络嵌入的工作[7~10]，虽然他们的效果都很优异，但是却都只关注有节点邻接关系定义的社区身份，而基于由节点拓扑结构定义的结构身份的研究截至撰稿时却依旧十分匮乏。2017 年，Ribeiro 等人提出了 struc2vec，一种从结构身份学习节点表征的新方法，其中他们把结构身份定义为基于网络结构和节点关系的对称性概念 [11]。如图 1 所示，图中 A1-E1 和 A2-E2 尽管并不相邻，甚至相隔一个复杂网络，但它们分别拥有相似的结构，因此也应当具有相似的结构身份。

更早针对网络结构身份的研究往往是根据邻接关系等信息进行相似度计算，如 Glen Jeh 等人提出 SimRank 算法 [12]，或者归于预先设定的若干类别中，而不能进行高维空间中的结构特征提取，较为接近的则是 Kleinberg 提出的 HITS 算法[13]，将互联网的网页根据结构特征分为枢纽页面(Hub)和权威页面(Authority)，这其实可以看做一种简单的结构身份表征，但只能在两个维度中进行，具有很大的局限性。而该细分领域最前沿的成果之一则是 Ribeiro 等人提出的 struc2vec[11]，它首先根据节点每一层邻居的度序列利用 DTW(Dynamic Time Warping，动态时间变换)算法两两计算节点间的结构相似度，然后据此建立一个分层网络，网络每

收稿日期：2018-09-15；**修回日期：**2018-10-29　**基金项目：**国家重点研发计划重点专项项目（2018YFB1004702）；国家自然科学基金资助项目（61822206，61532012，61602303，61829201）CCF-腾讯犀牛鸟项目（20180116）；天津先进网络重点实验室开放课题

作者简介：孟亚文（1993-），男，河南驻马店人，硕士研究生，主要研究方向为数据挖掘、网络嵌入和数据可视化研究；傅洛伊，女，特别副研究员，主要研究方向为社交网络、数据挖掘工程等；王新兵，男（通信作者），教授，主要研究方向为无线通信、社交网络、知识图谱等（xwang8@sjtu.edu.cn）.

一层都代表 N 阶邻居内的相似性，层内是完全图，每条边的权重代表节点在某个半径下的相似度，层间同一个节点对应的节点相互连接，然后在这个新的多层相似性权值网络中利用流行的随机游走+词向量嵌入模型的方法进行网络嵌入，最终获得每个节点结构身份的特征表示。该方法确实能很好的学习到网络节点的结构身份，但其相似性网络的构建过程却十分复杂，需要预先构建每个节点不同半径下邻居的度序列，而后两两计算每个半径下的相似度，最后才能进行网络嵌入，在实际计算中，度序列和相似度的计算通常占总时间开销的 90%以上，算法效率较低。



图 1　结构身份的演示。图中节点 A1、A2 拥有相似的结构，因此结构身份也应该相似。

Fig. 1　Demostration of structure identity, node A1 and A2 should have similar structure identities due to their similar local structures.

因此，本文提出了一种全新的算法，利用递归投影的方法进行结构身份的网络嵌入，并在可视化结果、分类精度、对抗随机边采样的鲁棒性和算法效率上与最前沿的算法进行比较。

1　基于递归投影的结构性网络嵌入

1.1　结构性网络嵌入的要求

为了并保证结构性网络嵌入的效果，学界认为良好的结构性网络嵌入算法应当至少满足以下三种性质：

a)结构相似的节点，其网络嵌入结果也应当相似，如图 2 中的 A1 与 A2 节点，这保证了算法的平滑性和稳定性，相似的输入对应的输出也是相似的。

b)结构不同的节点，其网络嵌入结果也尽量不同，如图 2 中的 A1 与 B 节点，这保证了算法的区分度，能够尽可能地区分具有不同结构的网络节点。

c)网络嵌入结果与节点属性、空间距离等无关，如图 1 中 A1 与 A2 节点，显然一个网络节点的拓扑结构特征只与它的局部邻接节点有关，因此位于网络中不同位置的节点，若其结构相似，其嵌入也应该相似。

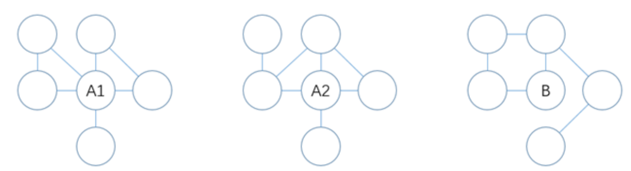


图 2　具有不同结构身份的网络节点演示

Fig. 2　Demostration of nodes with different structure identities

1.2　结构性网络嵌入的要求

根据 1.1 节中的要求，本文提出了一种基于递归投影的结构性网络嵌入算法，即 RSNE(recurrent structural network embedding)，算法核心思想在于将每个节点的结构身份视为其直接邻居结构身份的非线性投影，也即是其结构特征向量由其邻居的结构特征向量经由非线性映射迭代而得到。算法分为三个步骤，依次为基于直接邻居的结构身份初始化、基于非线性投影的迭代计算、基于数据标准化的后处理，以此满足上述三条性质，并生成合理的结构身份特征表示，下文将具体阐释本算法能够满足上述性质的原因，并设计实验予以验证。

1.2.1 基于直接邻居的结构身份初始化

考虑一个无向、无权重网络 $G = (V, E)$，其中 V 为点集，E 为边集，如图 2 中 $A1$、$A2$、B 三点，本文会发现结构身份相似的节点，其本身的度必然是相似的，其直接邻居也必然是相似，后者的相似性同样可以由度来衡量。此外，由度带来的结构身份差异应当是负指数型分布的，例如度为 100 和 101 的点，其结构差异显然小于度为 5 和 6 的点。

基于以上考虑，本文提出了如下的结构身份初始化方法：

为从 1 开始的每个度 d_i, $i \geqslant 1$,分配一个单位随机向量 u_i,

对每个节点 $v \in V$，度为 d_v，则其向量初始化为

再将每个节点直接邻居的初始向量投影到该节点上，即

其中第 1 步保证由度带来的结构身份差异为负指数型分布，第 2、3 步分别将节点本身和直接邻居的基本结构特征纳入初始特征向量中，为后续计算打好基础。

1.2.2 基于非线性投影的迭代计算

初始化的结构特征向量已经能够捕捉每个节点的基本结构信息，但是无法将更大尺度的拓扑结构纳入考虑，因此，本文继续通过非线性投影迭代的方法，即是将节点邻居的特征向量非线性映射为新特征后投影成当前节点的特征向量，算法描述如下：

对每个 $v \in V$，令

其中:令 D 为特征向量的维度，有

考虑邻接矩阵为 N_{n*n}，n 为图中节点数量，所有节点的特征向量组成特征矩阵 F_{n*d}，则算法的每次迭代都可表示为

1.2.3 基于数据标准化的后处理

经过递归投影后，特征矩阵 F_{n*d} 的每一维分布都不相同，可能会影响结构性网络嵌入的效果，因此需要进行数据标准化，即把 F_{n*d} 的每一维的特征值都线性变换为均值为 0，标准差为 1 的标准化数据。

1.2.4 算法总结

算法第一步以一阶邻居作为特征得到的初始化特征向量能够描绘出网络节点最基本的结构特性，同时使用指数分布来保证算法的平滑性和区分度，为后续更精细的特征提取提供基准。

第二步的递归投影则是为了将邻居节点的特征向量融入到中心节点中，这样便使得中心节点的特征向量能够描绘出更大范围内的结构特性，同时递归方式也能保证更近的节点拥有更大的投影权重，使得算法结构更加合理。

第三步的数据标准化则是为了保证特征向量的有效性，避免向量中某些随机产生的极端维度造成意外干扰。

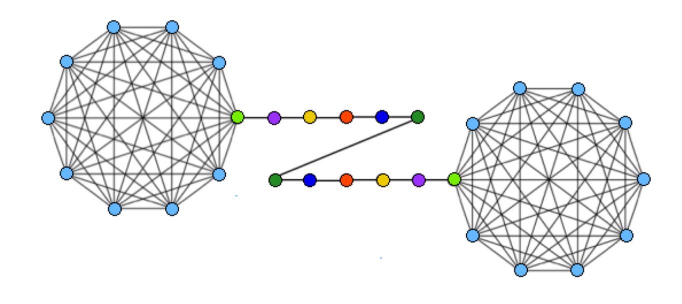
下文将进行多项实验以全面测试并证明本算法能够满足 2.1 中所提出的要求，并获得优秀的基于结构特征的网络嵌入结果。

2 实验设计与分析

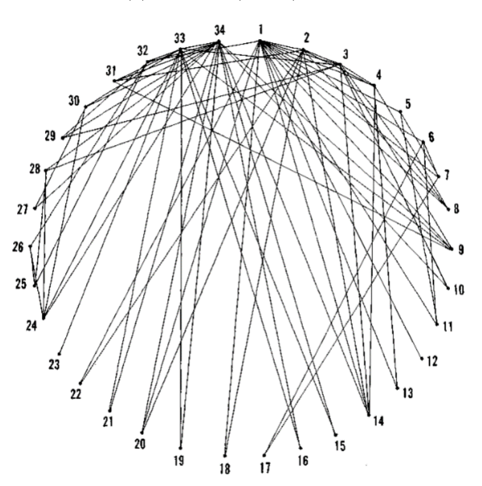
本节将参考 struc2vec[11]进行多项实验的分析和对比，并从可视化结果、结构特征向量与节点间相似度的相关性、对抗随机边采样的鲁棒性、针对真实数据集的分类精度以及算法效率等方面衡量算法效果，并与最前沿的算法，也即 2017 年的 struc2vec 算法[11]进行对比。

2.1 杠铃图

图 3(1)是一张典型的杠铃图，可表示为 $B(10,10)$，含义为左右两侧各有 10 个点为全连接，中间 10 点依次连接沟通两侧。显然图中颜色相同的节点拥有完全相同的结构特征，如左右两侧浅蓝色的 18 个点。



(a)杠铃图 $B(10,10)$ [11]



(b) 空手道俱乐部关系图 [14]

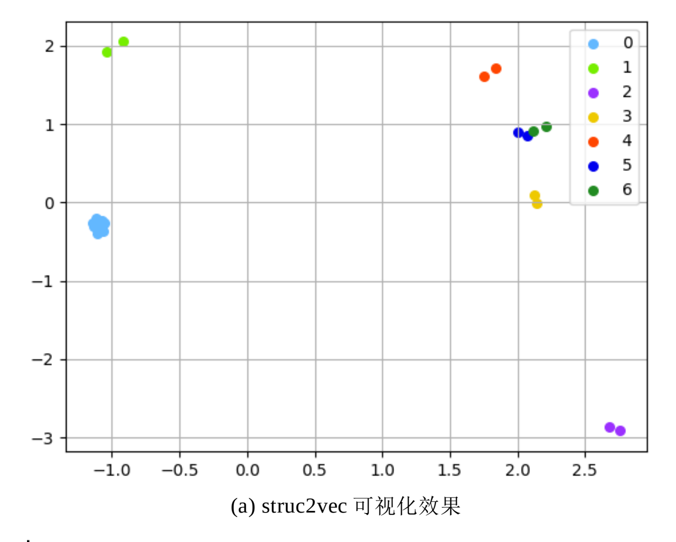
图 3 两种特殊的图

Fig. 3 Two special graphs

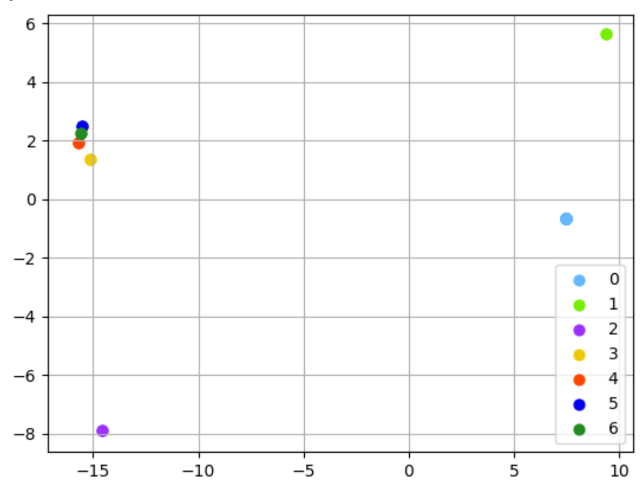
图 4 展示了本文 RSNE 与 struc2vec 算法在 128 维特征空间的学习结果经过 PCA 降维后在平面的可视化效果。可以看出，两者都能把共计 7 种点很好地区分开，其中结构特征明显的 0-2 类各自分散得很开，而处在图中间，结构特征类似的 3-6 类的点则较为接近。两者不同之处在于 struc2vec 同一类之间也会有一定距离，这是随机游走和词向量模型训练过程中引入的不可避免的随机性，甚至到了类 5 和类 6 的混叠，而 RSNE 则没有这个问题。

2.2 空手道俱乐部镜像关系图

空手道俱乐部关系图由 34 个点、78 条边组成，是网络嵌入研究中常用的一张图，如图 3(b)所示。这里仿照 struc2vec，将该图点、边复制一遍后仍放入原图，构成一个有 34 对节点的镜像图，下面分别使用 RSNE 和 struc2vec 在该镜像图中运行，并各自得到 128 维结构特征向量，并将结果数据标准化后计算对应镜像节点对的平均欧氏距离和所有节点对的平均欧氏距离，同时计算后者和利用 DTW 算法得到节点相似度分析 Pearson 相关系数，其结果如表 1 所示。



(a) struc2vec 可视化效果



(b) RSNE 可视化效果

图 4 杠铃图效果对比

Fig. 4 Comparison on the barbell graph

表 1 空手道俱乐部镜像关系图的结构性网络嵌入结果

Table 1 Structural network embedding results of the mirrored Karate Club graph

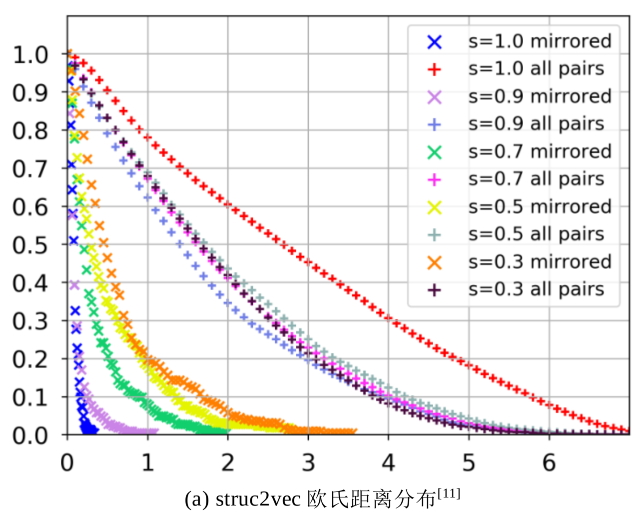
方法	镜像对距离	所有点距离	Pearson-0 层	Pearson-2 层	Pearson-4 层
Struc2vec	3.07	15.54	0.83	0.71	0.70
RSNE	0.00	11.77	0.93	0.66	0.58

可以发现，在欧氏距离上，本文提出的 RSNE 可以保证镜像对距离为 0，这比收到随机性影响的 struc2vec 要优秀得多。同时在节点距离和 DTW 相似度的相关性上，RSNE 在浅层网络结构的相关性更好，而在深层网络结构的相关性更差，这事实上是非常合理的，因为 DTW 对节点每层邻居的计算都是同权的，但显然浅层的邻居对节点的结构身份影响更大。

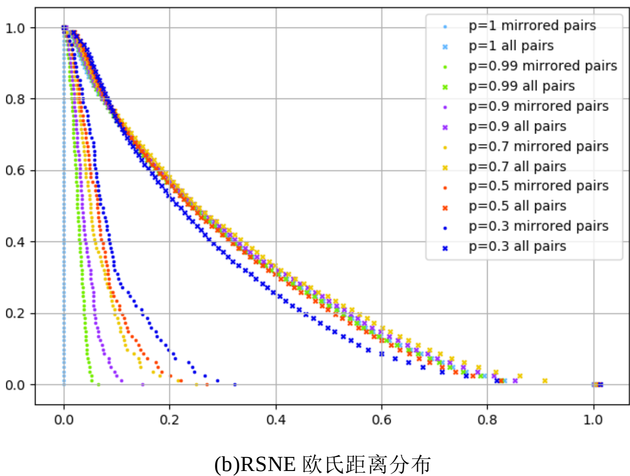
2.3 对抗随机边采样

本实验使用 Facebook 的一份社交网络图，其中包含 224 个节点，3192 条边，记为 $G = (V, E)$，而后将 G 中每条边 $e \in E$，按照概率 s 随机采样保留，得到 $G^* = (V, E^*)$，将 G, G^* 放入同一个网络中进行结构性网络嵌入，而后计算

G, G^*中对应点的欧式距离，该距离分布可以表示如图 5 所示。



(a) struc2vec 欧氏距离分布[11]



(b)RSNE 欧氏距离分布

图 5 边采样模型下对应点欧氏距离的分布

Fig. 5 Distance distribution between node pairs under the edge sampling model

显而易见，随着采样概率的降低，G^*中保留的边逐渐减少，节点对之间结构身份差异也将越来越大，因而其结构特征向量的欧氏距离也应该越来越大，图 5 的实验结果与此完全一致，边保留较多时节点对的特征向量差距不大，甚至在丢失大约 70%的边时，算法仍能维持节点对之间结构向量的相似性，同时所有节点对的平均距离仍保持在较大的水平，这证明 RSNE 在鲁棒性上达到了和前沿成果相当的表现。

2.4 针对机场数据的分类实验

网络嵌入的应用之一就是节点分类，本文采用 struc2vec[11]提供的美国和欧洲的机场航班往来网络进行分类实验，以美国机场网络为例，图中共有 1190 个节点，代表各个机场，共有 13599 条边，代表航班往来。所有机场按照人流量和航班数等信息分为 4 类，本实验首先利用 RSNE 和 struc2vec 进行结构性网络嵌入，在结果中选取 80%的样本作为训练集，剩余样本作为测试集，通过逻辑回归方法进行分类，分类结果如表 2 所示。

可以看出 RSNE 的分类效果与 struc2vec 分类准确率十分接近，但是其算法复杂度和实际时间开销却远小于 struc2vec，具体可见 2.5 小节。

2.5 算法时间开销

针对以上实验所涉及的五张图，本文分别统计了 RSNE 和 struc2vec 的时间开销，结果记录为表 3，可以看到，对同样的网络进行结构特征的学习，RSNE 所消耗的时间远远小于 struc2vec，甚至不到后者的 1%，但在鲁棒性、分类精度上到了相似的结果，这充分说明的本方法的有效性和优越性。

表 2 美国和欧洲的机场航班往来网络分类准确率

Table 2 Classification accuracy of US and European airport flight network

机场	美国	欧洲
Struc2vec	0.64	0.59
RSNE	0.63	0.61

表 3 在不同图上，RSNE 和 struc2vec 的时间开销

Table 3 Time consumption of RSNE and struc2vec on different graphs

图名	点数	边数	Struc2vec 耗时	RSNE 耗时
杠铃图	30	101	18.78	0.06
空手道俱乐部图	68	155	21.82	0.10
Facebook 图	224	3192	33.09	0.30
欧洲机场图	399	5995	52.88	0.50
美国机场图	1190	13599	195.12	1.83

3 结束语

本文将结构身份定义为节点所在局部网络的拓扑结构特征，它通常能代表节点在网络中的作用、地位等，如论文引用网络的中高质量论文、社交网络中的明星用户。本文针对当前网络嵌入在结构身份方面的匮乏和前沿成果的不足，提出了旨在学习网络中节点结构身份在向量空间中特征表示的新颖而精简的算法，RSNE。

RSNE 首先进行基于度和直接邻居的特征向量初始化，而后利用递归投影技术学习节点更深层、更精细的结构特征，最后通过数据标准化保证输出结果的可用性。由于本方法未使用两两计算相似度的方式进行网络结构身份的采样，因而能够将算法复杂度控制在较低水平。

在多个图上进行的实验表明，RSNE 能够有效捕捉网络节点的结构身份，在保证分类准确性的同时，也能够维持良好的鲁棒性，而且在计算过程中花费的时间远远小于相关前沿成果，如 struc2vec。

本文认为在更注重节点结构的网络数据挖掘任务中，如某学术会议中学术新星的发掘，结构性网络嵌入方法拥有 node2vec、LINE、Deepwalk 等方法不可比拟的优势，这也是本文提出 RSNE 的贡献和意义所在，同时本文所用到的简化网络嵌入的思想也可以应用到更为广阔的领域中，如应用到室内定位网络中等[15]。

参考文献：

[1] Cui Peng, Wang Xiao, Pei Jian, *et al.* A survey on network embedding [J]. arXiv preprint arXiv: 1711. 08752, 2017.

[2] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations [C]//Proc of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.NewYork: ACM Press, 2014: 701-710.

[3] Mikolov T, Chen K, Corrado G, *et al.* word2vec [J]. Google Scholar, 2014.

[4] Tang Jian,Qu Meng,Wang Mingzhe,*et al.* Line:large-scale information network embedding[C]//Proc of the 24th International Conference on World Wide Web.2015: 1067-1077.

[5] Grover A, Leskovec J. node2vec:scalable feature learning for networks [C]//Proc of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.New York: ACM Press, 2016: 855-864.

[6] Qiu Jiezhong, Dong Yuxiao, Ma Hao, *et al.* Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec [C]// Proc of the 11th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2018: 459-467.

[7] Bojchevski A, Shchur O, Zügner D, *et al.* NetGAN: Generating Graphs via Random Walks [J]. arXiv preprint arXiv: 1803. 00816, 2018.

[8] Dai Quanyu, Li Qiang, Tang Jian, *et al.* Adversarial network embedding [J]. arXiv preprint arXiv: 1711. 07838, 2017.

[9] Feng Rui, Yang Yang, Hu Wenjie, *et al.* Representation learning for scale-free networks [J]. arXiv preprint arXiv: 1711. 10755, 2017.

[10] Wang Hongwei, Wang Jia, Wang Jialin, *et al.* GraphGAN: graph representation learning with generative adversarial nets [J]. arXiv preprint arXiv: 1711. 08267, 2017.

[11] Ribeiro L F R, Saverese P H P, Figueiredo D R. struc2vec: learning node representations from structural identity [C]//Proc of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press,2017: 385-394.

[12] Jeh G, Widom J. SimRank: a measure of structural-context similarity [C]// Proc of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM Press, 2002: 538-543.

[13] Kleinberg J M. Authoritative sources in a hyperlinked environment [J]. Journal of the ACM, 1999, 46(5): 604-632.

[14] Zachary W W. An information flow model for conflict and fission in small groups [J]. Journal of Anthropological Research, 1977, 33(4): 452-473.

[15] Shangguan Longfei, Yang Zheng, Liu A X, et al.STPP: spatial-temporal phase profiling-based method for relative RFID tag localization[J]. IEEE/ACM Trans on Networking, 2017, 25(1): 596-609.